Vol. 34 No. 4 情報処理学会論文誌 Apr. 1993

複合語の構造化に基づく対訳辞書の単語結合型辞書引き

宮崎正弘† 池原悟†† 横尾昭男††

機械翻訳システムにおいて，大規模辞書を効率的に構築し，維持管理することはきわめて重要である．本論文では，機械翻訳システムの解析辞書（日本語辞書）と変換辞書（対訳辞書）にどのような基準，条件で見出し語を収録したらよいかについて論じ，各種辞書の単語収録単位の違いを吸収するものとして単語結合型辞書引きを提案した．複合語は短単位語（語基）を組み合わせて数限りなく生成される．従って，このような複合語は，解析辞書には原則として収録せず，複合語は語基の組合せとして，その内部構造を解析する．一方，変換辞書には目的言語に応じて適切な訳を生成するため語基のほかに複合語を収録し，複合語の内部構造の解析により生成された部分複合語を基に，複合語内の語基を組み合わせて変換辞書引きを行う．この過程により，複合語は変換辞書にある見出し語の最適な組合せに再構成される．本方法により，数限りなく生成される複合語を原則として解析辞書に収録する必要がないので解析辞書のコンパクト化が図れ，解析辞書と変換辞書の独立性を確保でき，大規模辞書の効率的な構築・維持管理が可能になると共に，日本語の複合語に対する解析と変換処理の調和が実現できた．

Combined Word Retrieval for Bilingual Dictionary Based on the Analysis of Cmpound Words

Masahiro Miyazaki,† Satoru Ikehara†† and Akio Yokoo††

Dictionary construction method for machine translation supported by combined word retrieval is proposed. In this method, translation dictionaries are comprised of two kinds of dictionaries: "analysis dictionary" for analyzing original language which contains information only for morpheme and "transfer dictionary" for translations of words which contain information of compound words as well as of morphers. Combined word retrieval is applied for the latter dictionary to retrieve compound words. "Analysis dictionary" becomes compact independently from transfer dictionary by this method and high quality translation of compound words can consistently be performed. Large-scale dictionaries for machine translation system can be developed consistently and also maintenance becomes easier.

1. まえがき

日本語文を目的言語に翻訳する場合，短単位語（語基☆）を組み合わせて適切な訳を生成できない長単位語（複合語☆☆）が多い．特に現在，機械翻訳の対象となっている科学技術文書（論文，マニュアルなど）や新聞記事文（産業，経済分野など）などでは，専門用語を中心に多くの複合語が出現する．

† 新潟大学工学部情報工学科
Department of Information Engineering, Faculty of Engineering, Niigata University
†† NTT 情報通信網研究所
NTT Network Information Systems Laboratories

☆ 単独で語を構成できるか否かによって語基と接辞に分けられるが，ここでは，接辞も語基に含め，以下単に語基と呼ぶ．

☆☆ 上記で定義した語基を複数個結合して作られる語を合成語と呼ぶこともあるが，ここでは"複合語"と呼ぶことにする．

従来，多くの機械翻訳システムでは，このような複合語を解析辞書に収録し，入力文の解析において複合語をそれ以上分割できない一つの単語とみなし，複合語の内部構造は解析していない．しかし，このような方式では，以下のように，辞書作成過程のみならず，アルゴリズム構成上も重要な問題が生じる．

①複合語は無限に生成されるので，すべての複合語をあらかじめ辞書に収録しておけない．

②辞書の収録語数が際限なく増大し，辞書の構築，維持・管理がやりにくくなる．特に，解析に必要な意味属性などの意味情報を，語基だけでなく，多くの複合語に対して付与するには，多大の工数を要する．

③登録語数の増大は，分かち書きにおける単語分割などの曖昧性を増加させ，これを解消するために有効な曖昧性解消機構が必要となるが，この機構が不十分な場合には，解析精度の低下を招く（図1の例1

を参照).

④複合語外かから複合語内の語基への係り受け解析ができないため，複合語の周辺の語との意味的関連を考慮した柔軟な訳出ができない（図1の例2を参照).

これに対して，本論文では，機械翻訳システムにおいて，上記問題点を解決するため，解析辞書（日本語辞書)，変換辞書（対訳辞書）にどのような基準，条件で見出し語を収録したらよいかについて論じ，各辞書の単語収録単位の違いを吸収する方法として単語結合型辞書引きを提案する.

本方法では，解析辞書，変換辞書の独立性を確保するため，解析辞書には原則として語基を収録し，変換辞書には目的言語に応じて適切な訳を生成するため，語基のほかに複合語も収録する．まず，日本語の複合語は語基の組合せとして複合語の内部構造を解析し，同時に複合語外から複合語内の語基への係り受け等も解析する．以上の解析結果に基づき，複合語内の語基を組み合わせて変換辞書引きを行う単語結合辞書引きを行い，複合語を変換辞書にある見出し語の最適な組合せに再構成する.

本方式は，日本語の用言連用形が副詞化したり（例：激しく→violently)，日本語の名詞＋格助詞「の」が形容詞化するなど（例：近代の→modern)，日本語と目的言語の単語や品詞が1対1に対応しないため，解析辞書と変換辞書の単語収録単位，品詞の違いを吸収するためにも有効である.

2. 日本語辞書と変換辞書の単語収録条件

機械翻訳用の辞書において，解析辞書と変換辞書の独立性を確保することは，きわめて重要である．そのために解析辞書は原言語の解析を行うのに必要な情報のみを収録し，目的言語に依存した情報は原言語と目的言語の対応を記述した変換辞書に収録することが望ましい．このような辞書構成をとることにより，多言語翻訳を行う場合でも，目的言語に対応した変換辞書を準備しさえすれば，解析辞書は何の変更も加えることなく共通に使える．また，変換辞書の変更によって，解析が大きな影響を受けることがないため，解析過程が安定する.

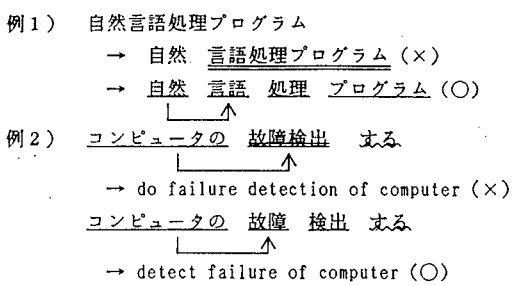


図1 複合語の辞書収録による処理誤りの例
Fig. 1 Analytical error by registration of compound word in a dictionary.

ここでは，解析辞書としての日本語辞書と変換辞書の独立性の維持を狙い，複合語，活用語に対する辞書収録の形態と正書法の確立されていない日本語における表記のゆれの吸収の方法，等について論じる．また，これらの議論を通じて，解析辞書，変換辞書の単語収録条件を明らかにする.

2.1 複合語の扱い

日本語では，造語力の強い漢字等により助詞を介さずに名詞（接辞などの名詞相当語を含む）が二語以上連続した名詞連続複合語（以下，単に“複合名詞”と呼ぶ）が限りなく作り出される．このような複合名詞は科学技術文書や新聞記事といった，現在，機械翻訳の対象となっている実用文には数多く現れる．しかし，このような複合名詞を，あらかじめすべて辞書に収録することはできない.

そこで，辞書の見出し語は語基を原則とし，複合語は処理によって語基の組合せに分割して扱う．ただし，語基の組合せに分割できない複合語は長単位で辞書に収録する．専門語を初めとする複合語の多くは，語基の組合せに分割できるため，長単位で解析辞書に収録すべき単語は多くないのに対して，語基を組み合わせて適切な訳出ができない複合語は，かなり多い．従って，変換辞書には，このような複合語を多数，収録する必要がある.

ここで，4章で提案する変換辞書における単語結合型辞書引きを用いれば，解析辞書と変換辞書の単語収録単位の違いを吸収できる点に着目して，解析辞書には，変換辞書に収録した複合語を収録しない．このようにすることにより，解析辞書と変換辞書の独立性を確保できるだけでなく，解析辞書をコンパクトにすることができ，辞書の構築や維持管理がしやすくなる．特に，解析に必要な意味属性などの意味情報を多くの複合語に対して付与しなくてすむ等の利点がある．また，複合語の構造解析アルゴリズムを強化しさえすれば，複合語を多数，解析辞書に収録することによる単語分割誤りを減らすことも期待できる.

2.2 活用語の扱い

単語の区切りがなく，べた書きされる日本語文の自

動分かち書きを効率的に行うため，日本語の活用語は，通常，以下の二つの形態で解析辞書に収録される．

①規則的な活用を行うもの

……不変化部分と変化部分に分離して収録する．

②不規則な活用を行うもの（助動詞，変格活用動詞など）……すべての活用形を収録する*．

これに対して，変換辞書には活用語の代表形（通常，終止形）を収録し，解析辞書と変換辞書の単語収録単位の違いを吸収するため，活用語を終止形に変換し，元の活用形，法情報（活用語が助動詞，補助用言で様相，相・時制などの属性をもつ場合）を付与する．ここで，活用語が助動詞などで，活用語に対応する訳語が目的言語にない場合には，法情報を基にした構文構造の変換等を行う必要がある．

2.3 表記のゆれの扱い

日本語では明確な正書法が確立されていないため，同じ語を漢字，ひらがな，およびその混ぜ書きといった各種の形式で表記したり，数詞を漢数字，算用数字，およびその混ぜ書きといった各種の形式で表記したり，送りがなのゆれがあるなど，表記上のゆれがある．このような表記上のゆれは，日本語内に閉じたものであるので，日本語文の解析処理の早い段階で解消しておき，変換辞書には，表記のゆれのある語のうち，その代表形（以下，これを標準表記と呼ぶ）のみ見出し語として収録することが，変換辞書のコンパクト化の観点から望ましい．

日本語の表記上のゆれを解消するため，表記のゆれのある語について，そのすべての形を解析辞書に見出し語として収録し，見出し語に対応する標準表記を辞書中に記述することが考えられる．しかし，このような方法では解析辞書の収録語数が膨大となってしまう．そこで，以下のような辞書と処理を併用した方法で表記のゆれを解消する方法を提案する．

（1） 処理による表記のゆれの解消

① 漢字の異字体の解消

入力文中にある漢字の異字体は，異字体と代表字体の対応を示すテーブルを用いて，代表字体に変換する．

例）附属 → 付属，廻転 → 回転

② カタカナ外来語の表記のゆれの解消

入力文中にある外来語表記用の特殊なカタカナ文字（α）は，α と代表字体の対応を示すテーブルを用いて代表字体に変換する．

例）ヴァイオリン → バイオリン

また，分かち書き処理を行った場合，カタカナ既知語*（β）の直後にある長音「ー」一文字の未知語は，直前の β と統合して一語（標準表記は β のまま）とする．

例）データ／ー → データー（「／」は単語境界）

③ 繰り返し記号に関するゆれの扱い

入力文中にある繰り返し記号（「々」，「ゝ」，「ゞ」）は，対応する文字に置換する**．

例）たゝみ → たたみ，かゞみ → かがみ

いちゝゝ → いちいち，一歩々々 → 一歩一歩

ただし，「々」が連続しない場合（例：「佐々木」「日々」「三々五々」）は，例外的に「々」を用いた形も解析辞書に収録しておき，「々」を含む単語が辞書にない場合のみ上記のような文字の置換を行う．これは，固有名詞などは，通常「々」を用いて表記すること，「々」を漢字で置換することにより入力文字の情報の一部が縮退すること（たとえば，副詞「一々」は「一一」となり，数詞と同型語となってしまう）などのためである．

④ 数詞の表記のゆれの解消

種々の形式で表記された日本語の数表現を文献1で提案されている日本語の数表記の標準形に変換する．

（2） 辞書による表記のゆれの解消

① 送りがなのゆれの解消

すべての可能な形を見出し語として辞書に収録する．

例）行(な)う→行う，行なう

申(し)込(み)→申込，申込み，申し込み

② 同一字種内の表記のゆれの解消

すげての可能な形を見出し語として辞書に収録する．

例）二葉 ←→ 双葉，衣装 ←→ 衣裳，

コンベア ←→ コンベヤ

③ 異なった字種間の表記のゆれの解消

すべての自立語（接辞，補助用言を含む）について，そのひらがな書き，混ぜ書き等の形を辞書に収録することは，辞書の収録語数の増大を招くだけでなく，分かち書き処理における単語分割誤りの要因ともなる．従って，自立語のうちひらがな書き，カタカナ書き，

---

* 「行っ」（五段動詞「行く」の音便形）のように特定の活用形で例外的な活用をするもの（カ行は通常イ音便となる）は，活用形で例外的な形のみ収録する．

* カタカナ語で長音のゆれについては，長音化しない方のみ，解析辞書に収録しておく．

** 解析辞書には，原則として繰り返し記号を用いない形で収録しておく．

混ぜ書きされることが多い語のみ日本語辞書に見出し語として収録する．

なお，机上シミュレーションによれば，解析辞書を通常の見出し語のほかに，読み（ひらがな表記）でも検索できるようにすれば，解析辞書の収録語数を増やすことなく，本来，ひらがな以外で表記されることが多いひらがな語の辞書引きができる．また，カタカナ→ひらがな変換した後，解析辞書を読みで検索することにより，本来，カタカナ以外で表記されることが多いカタカナ語の辞書引きができる．このような読みによる検索は，分かち書き処理において，ひらがな未知語，カタカナ未知語が検出された場合等に，必要に応じて行うことにより，分かち書き精度の向上が期待できる．

2.4 同義語の標準化

日本語内で，ある単語（γ'）を別の単語（γ）に置換しても意味の等価性が保たれる場合，解析辞書内にその代表形（γ）を記述することにより標準化し，変換辞書には代表形のみ記述することにより，変換辞書のコンパクト化が図れる．この場合，$\gamma' \rightarrow \gamma$ の置換によりその意味は変わらないが，γ' が特別な慣用表現や複合語を形成し，かつそのような表現において $\gamma' \rightarrow \gamma$ の置換不可の場合，γ を γ' の代表形とはせず，γ' を代表形とする．また，$\gamma' \rightarrow \gamma$ の変換では等価とならないが γ に適当な法情報（δ）を付与することにより等価となる場合，γ' は δ が付与された γ に変換される．

例）泳げる（γ'）→泳ぐ（γ）[δ='可能']

2.5 そ の 他

2.2〜2.4 節で述べた代表形，標準表記以外の語を変換辞書に収録しなければならない場合としては，2.1 節で述べた複合語のほかに，以下のようなものがある．

①日本語の用言の非終止形が，目的言語において他の品詞になる場合．

例）激しく（形容詞・連用形）
→ violently（副詞）

②日本語の用言＋付属語例が，目的言語において他の品詞になる場合．

例）注意して（動詞＋接続助詞）
→ carefully（副詞）

③日本語の体言＋付属語列が，目的言語において他の品詞になる場合．

例）木の（名詞＋格助詞：木材の意味）
→ wooden（形容詞）

④日本語において，用言とその格要素（名詞＋格助詞など）が，目的言語においてひとまとまりの単語，句などになる場合．

例）背が高い（名詞＋格助詞＋形容詞）
→ tall（形容詞）

⑤日本語において，格助詞＋用語＋付属語列（全体で格助詞相当語となる）が，目的言語においてひとまとまりの単語，句などになる場合．

例）に対して（格助詞＋動詞＋接続助詞）
→ to（前置詞）

上記①〜⑤については，解析辞書と変換辞書の単語収録単位や品詞の違いを吸収するため，単語結合型辞書引き，品詞変換などを行う必要がある．

3. 複合語の構造解析

複合名詞（サ変動詞，形容動詞の語幹となる複合語を含む）は，複合語のうちで最も出現頻度が高く，か

表 1 複合語における係り受け規則
Table 1 Dependency analysis rules for compound words.

番号	係り受けの型		例
1	前置助数詞－数詞		約 10、第 八 回
2	数詞－後置助数詞		二 本、50 パーセント
3	後置助数詞－助数詞承接型接辞		50Kg 強、数 % 台
4	数詞－数詞承接型接辞		100 未満、10 以下
5	固有名詞－固有名詞承接語	地名	東京 駅、関東 平野
		人名	平野 副 社長、一郎 君
		組織名	三井 信託 銀行
		その他の固有名詞	明治 時代、アイヌ 人
6	役職承接型接辞－役職		美濃部 前 都 知事
7	姓－名		加藤 二二三
8	包含関係のある地名の連接		神奈川 県 横須賀 市 武
9	接頭語－単語*1		極 超 短波
10	単語*1－接尾語		大型 機 用
11	非用言性名詞 －非用言性名詞		一月 一日 朝、県立 高校
12	単語*2－サ変動詞型名詞		データ 処理
13	サ変動詞型名詞－単語*2		処理 手順
14	単語*2－形容動詞型名詞		人気 絶頂
15	形容動詞型名詞－単語*2		特別 料金

単語*1：数詞、固有名詞、形式名詞を除く名詞
単語*2：名詞

つ解析上，問題となる．ここでは，このような複合名詞を対象に，変換辞書の単語結合型辞書引きを行うのに必要な構造解析法について述べる．

3.1 複合名詞内の係り受け解析

複合名詞の解析については，既に文献 2 で複合名詞を構成する単語間の意味的結合関係を係り受け解析によって解析する方法が提案はれており，それによっ

表 2　サ変動詞型名詞の係り受け規則
Table 2　Dependency analysis rules for verbal nouns.

項番	係り受けの型	係り受けの条件：係り先単語	係り受けの条件：備考	例
1	格関係 が、を、に、 γ_1 のために、 α ・・・・ 格関係or連体修飾or並列 が、を、に、 α_1 のために、 α ・・・・ する、 するという、 する事により生じた	αの前方の単語*1 αの直近の単語より順次、係り受けをチェックする。本チェックは項番2のチェックの前に行う。 α、α_1の一方が非動作性名詞としても使われる場合、または自動詞性の場合、当該単語をγ_1とする。上記の置換が行われず、α、α_1の一方が形容動詞化する場合、当該単語をβ_1とする。	[格係格] ・動作主、対象、起点、目標、道具、手段、材料、目的、結果、原因、理由、時間、場所、数量、様態、状態、共同 [格関係or連体修飾or並列] 以下の①または②の場合は並列とする。 ① αとα_1の意味属性が同等で、かつサ変動詞化した場合の必須格の格パターン*2が一致する。 ② 格関係、連体修飾の何れもが成立しない。	・情報 処理（を） ・データ 自動 収集（を） ・故障 検出（を） [格関係] ・処理 命令（で）（する） [連体修飾] ・整理 整頓（し） [並列]
2	連体修飾 α する γ_2 するという する事により生じた	αの後方の単語*1 αの直近の単語より順次、係り受けをチェックする。	[連体修飾] ①γ_2（が、を、に、のために、・・・） αする → αするγ_2 （γ_1とγ_2の格は重複しない） ②αとγ_1の意味属性が同等 →αするというγ_2（同格） ③γ_2の意味属性が抽象物、抽象的関係を表す特定の語 → αすることによって生じたγ_2	・予想 最高 気温（を）（する） ・発光 現象（するという） ・勤務 成績（する事によって生じた）

*1：接頭語＋単語は一つの単語とみなす。
*2：用言のとりうる格、格を表示する助詞、格要素となる名詞の意味属性。

（凡例）
α_1:サ変動詞型名詞、 β_1:形容動詞型名詞、 γ_1:その他の名詞

表 3　形容動詞型名詞の係り受け規則
Table 3　Rules of dependency analysis for adjective verbal noun.

項番	係り受けの型	係り受けの条件：係り先単語	係り受けの条件：備考	例
1	格関係 が、に、・・ γ_1 β 格関係 が、に、・・ α_1 β	βの直前の単語* 項番2のチェックの前に本係り受けをチェックする。	[格関係] ・主体／対象 ～がβだ ～が～にβだ ～が～をβだ ～は～がβだ	・栄養 豊富（が） ・研究 熱心（に）
2	連体修飾 な、的な、の β γ_2 連用修飾 に、的に β α_2 連用修飾or並列 に、的に、 β かつ β_2	βの後方の単語* βの直近の単語から順次、係り受けをチェックする。 本係り受けが成立せず、α_2が非動作名詞としても使われる場合には $\alpha_2 \rightarrow \gamma_2$ とする。 例）大型／計画	[連体修飾] ①γ_2（が、に、・・・）βだ → β（な、的な、の）γ_2 （γ_1とγ_2の格は重複しない） ②γ_2のγ_1がβだ → γ_1がβ（な、的な、の）γ_2 （γ_1がγ_2の部分、属性、所有物、動作である。） [連用修飾or並列] βが連用修飾的に使われない場合、並列とする。	・一様 データ（が） ・大型 切断 機（が） ・栄養 豊富 食品（の） ・高速 処理（に） ・自動 データ 処理（的に） ・完全 自動（に） [連用修飾] ・小型 軽量（かつ） [並列]

*：接頭語＋単語は一つの単語とみなす。

（凡例）α_1:サ変動詞型名詞、 β_1:形容動詞型名詞、 γ_1:その他の名詞

て高精度の複合語自動分割が実現されている．ここでは，文献 2 の方法に基づき複合名詞内の係り受け解析を行い，複合名詞を解析辞書にある語基の組合せに自動分割するが，以下のように係り受け規則を拡充した．従来の主要な係り受け規則を**表 1** に示す[2]．

（1） 用言性名詞の係り受け規則の拡充

サ変動詞化する名詞 α（サ変動詞型名詞），形容動詞化する名詞 β（形容動詞型名詞*）などの用言性名詞に関する係り受け規則を**表 2**，**表 3** のように拡張した．

主な拡張点は以下のとおりである．

①サ変動詞型名詞の係り受け規則

- α の前方の単語（直前の語以外）との格関係のチェック
- α と α が連接した場合における格関係，連体修飾，並列用法の判定
- α の連体修飾用法における同格のチェック

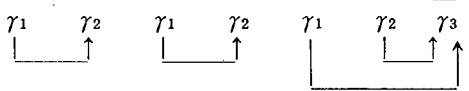
②形容動詞型名詞の係り受け規則

- β の直前の単語との格関係（主格以外）のチェック
- β の後方の単語との連用修飾，並列，連体修飾（主格以外）のチェック

（2） 非用言性名詞の係り受けの規則の拡充

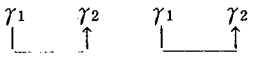
連接する非用言性名詞 γ_1 と γ_2 が意味的に強く結合するような γ_1 と γ_2 の一般名詞属性の組合せを共起属性表**の形で用意し，γ_1 と γ_2 の係り受けをチェックする．ただし，γ_2 の直後に「者」，「士」，「所」，「場」など語基との結合力の強い接尾語（強結合型接尾語と呼ぶ）γ_3 がある場合，γ_1 と γ_3 の係り受けをチェックし，γ_1 と γ_3 の係り受けが不成立の場合のみ，γ_1 と γ_2 のチェックを行う．

例）都市 住民，宇宙 基地，女性 弁護 士



なお，γ_1 と γ_2（または γ_3）の係り受けが成立せず，γ_1 と γ_2 の一般名詞属性が同等の場合には，γ_1 と γ_2 は並列とみなして係り受けを成立させる．

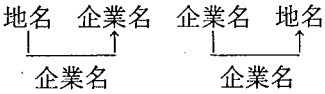
例）中学 高校，科学 技術



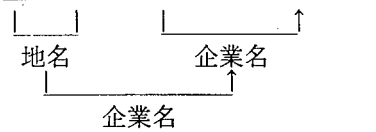
（3） 固有名詞の係り受け規則の拡充

①連接する固有名詞における特殊な共起パターンを検出し，係り受けを成立させる．

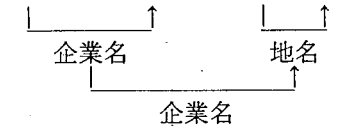
例）日本 IBM，東 芝 米国



東 北陸 富士通 ソフトウェア

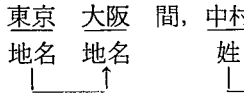


日電 ソフトウェア 東 北陸



②固有名詞が連接し，係り受けが成立しない場合，各固有名詞の固有名詞属性が同等ならば，各固有名詞は並列とみなして係り受けを成立させる．

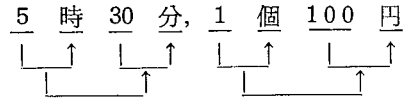
例）東京 大阪 間，中村 田中 両 氏



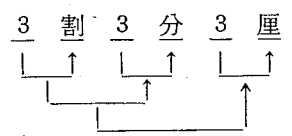
（4） 数詞の係り受け規則の拡充

①連接する数表現（数詞＋後置助数詞）における特殊な共起パターンを検出し，係り受けを成立させる．

例）5 時 30 分，1 個 100 円

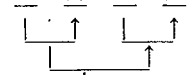


3 割 3 分 3 厘



②同型の数表現が連接し，係り受けが成立しない場合，各数表現は並列とみなして係り受けを成立させる．

例）1 日 2 日 の両日



3.2 複合名詞の構造解析

従来，複合名詞を構成する語基が意味的にどのように結合して，複合名詞を構成しているかを明らかにする複合名詞の構造解析については，十分研究されていない．複合名詞の意味的結合関係による分類に関しては，複合名詞を構成する語基の結合順序に着目して，

---

* 「自動」「～型」「～性」「近代」のような連体詞的に使われる連体詞型名詞を含む．

** 本論文で用いている一般名詞の階層的意味属性体系（一般名詞属性数＝約 2800）の最上位にある 68 の一般名詞属性の集合 $\{r_k\}$ から任意の二つの一般名詞属性（r_i および r_j）を選んで作成されるすべての一般名詞属性対 (r_i, r_j) に対し，「r_i の r_j」が意味的に成立するか否かチェックし，「r_i の r_j」が意味的に成立する一般名詞属性対をすべて集めて，共起属性表とした．

結合順序をパターン化したもの[3]，複合名詞を構成する語基に適当な付属語などを補うことにより語基間の意味的結合関係を明確化できることに着目して，4文字漢字列を分類したもの[4]など種々の提案がなされている．しかし，これらのものでは，複合名詞の分割や複合名詞を構成する語基間の意味的結合関係の分析は人間が行っており，これらをそのまま機械処理として実現できない．これに対して，機械処理を目指したものとしては，漢字複合名詞の構造を文脈自由文法（CFG）によってモデル化し，CFGの構文解析アルゴリズムによりその確率的構造解析を行ったもの[5]，漢字複合名詞を対象にサ変動詞型名詞の格フレームやサ変動詞型名詞の前後にくる名詞の意味属性・字面などを用いてその構造解析を行ったもの[6]，複合名詞内の名詞間の関係（並列・同格・修飾など）に着目した複合名詞の構造解析法[7]などがあるが，固有名詞まで含めた一般的な複合名詞に対し，統語情報だけでなく意味情報をも用いた構造解析法は提案されていない．これに対して，ここでは複合名詞内の係り受け解析結果を基に，補強CFG流の部分複合語生成ルールを用いた，固有名詞まで含めた一般的複合名詞に対する統語解析・意味解析融合型の構造解析法を提案する．

複合語を構成する語基数nが3以上の場合，複合語内の語基が部分的に結合してまとまり，複合語内により小さな複合語（部分複合語）を作る☆．このような部分複合語は，基本的には以下の過程により生成される．

部分複合語←語基＋語基

部分複合語←部分複合語＋語基|語基＋部分複合語

部分複合語←部分複合語＋部分複合語

すなわち，語基同士が結合して部分複合語を作り，このようにして作られた部分複合語に語基が結合したり，部分複合語同士が結合して，より大きな部分複合語が作られる．このように隣接した語基や部分複合語が結合する過程を繰り返して，より大きな構造を作りながら複合語が構成されると考えるのである．

ここでは，3.1節で述べた複合名詞内の係り受け解析の結果を基に，複合名詞内の語基がどのように結合して，部分複合語，より大きな部分複合語，さらに全体の複合語を構成しているのかを解析する方法について述べる．

☆ ここでは，最大の部分複合語を複合語と考え，$n=2$の場合も含んだすべての複合語を対象に議論を進める．

（1） 部分複合語生成ルール

部分複合語生成ルールを補強文脈自由文法流の書換え規則で以下のように表現する．

$X_i\{\mathrm{syn}=\alpha[,\mathrm{semg}=(\beta_1,\cdots),\mathrm{semp}=(\gamma_1,\cdots)]\}$

$\leftarrow y_j\{a_j\}z_k\{a_k\}\mid Y_j\{A_j\}z_k\{a_k\}\mid$

$y_j\{a_j\}Z_k\{A_k\}\mid Y_j\{A_j\}Z_k\{A_k\}$

$\{a_{j(k)}\}=\{\mathrm{syn}=\alpha'\ [,\mathrm{semg}=(\beta_1',\cdots),$

$\mathrm{semp}=(\gamma_1',\cdots),''\delta'']\}$

$\{A_{j(k)}\}=\{\mathrm{syn}=\alpha''[,\mathrm{semg}=(\beta_1'',\cdots),$

$\mathrm{semp}=(\gamma_1'',\cdots),''\delta'']\}$

X_i　：新たに生成される部分複合語

Y_j, Z_k：新しい部分複合語を構成する部分複合語

y_j, z_k　：新しい部分複合語を構成する語基

{ }内はX_i, Y_j, Z_k, y_j, z_kの品詞，一般名詞属性，固有名詞属性，表記を記述する．[ ]内は省略可であることを示す．

syn　：＝以下に品詞を表示

semg：＝以下に一般名詞属性のリストを表示

semp：＝以下に固有名詞属性のリストを表示

$''\delta''$　：語基または部分複合語の表記(δ)を" "内に表示

（2） 部分複合語の生成

数表現や固有名詞表現は隣り合った語基や部分複合語同士が局所的に次々に結合して，数詞や固有名詞相当の部分複合語を構成することが多い．一方，接辞や用言性名詞は数表現，固有名詞，その他の非用言性名詞，およびそれらから生成された部分複合語と結合して，さらに大きい部分複合語を構成することが多い．以上の点を考慮して，部分複合語生成ルールに以下のような優先順位をつけ，隣接し，かつ係り受けが成立する語基や部分複合語に対して，優先順位の高い順に本ルールを適用する．以上の過程によりボトムアップに部分複合語を生成していき，複合語に一致する部分複合語が生成された時点で生成を終了する☆．

[部分複合語生成ルールの優先順位]

①数詞関連ルール

②固有名詞関連ルール

1） 人名（役職関連表現を除く）関連

☆ 複合語に一致する部分複合語が生成されなかった場合，複合語αは最も大きな部分構造（部分複合語や語基）の組合せ$\alpha_1, \alpha_2, \cdots \alpha_N$に分割されている．$\alpha_1 \cdots \alpha_{N-1}$の中で係り先がないものについて係り受け規則を再度適用して，係り先を確定する．この結果，再度部分複合語生成規則を適用し，より大きな部分複合語の生成が可能となる．

2) 地名関連
3) 人名，地名，組織名以外の固有名詞関連
3) 組織名関連*
5) 人名（役職関連表現**）関連

③非用言性名詞関連

④接辞関連

1) 接頭語関連（接頭語が後方の語基に係る場合で，この時点でルールが適用できない場合，後方の語基から部分複合語にまとめられ適用可能となった時に適用する）
2) 接尾語（強結合型接尾語）関連

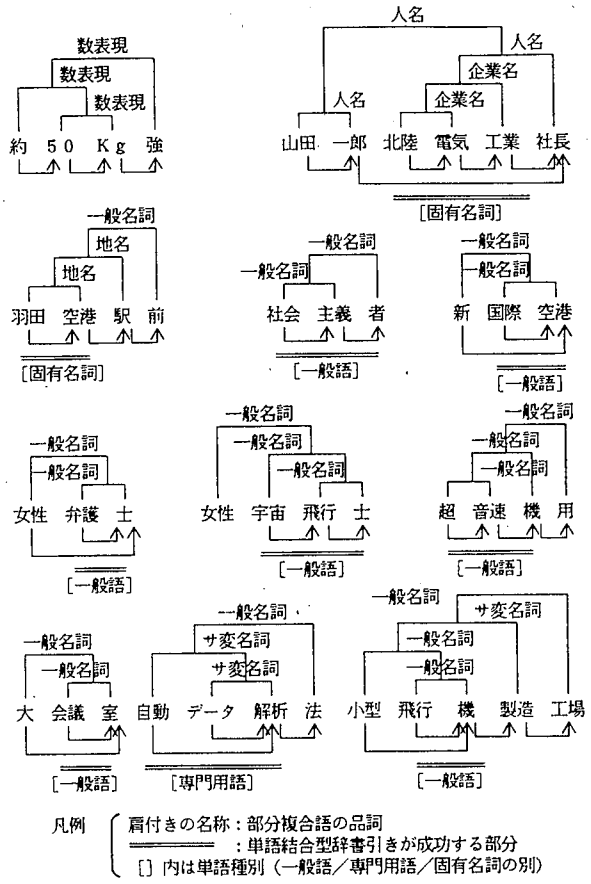


図 2 部分複合語の生成と単語結合型辞書引きの例
Fig. 2 Generation of partial compound word and combined word retrieval.

3) 接尾語（強結合型接尾語以外）関連***

⑤用言性名詞関連

部分複合語の生成例を**図 2** に示す．

3.3 複合語外から複合名詞内への係り受け解析

名詞句中に複合名詞がある場合，複合語外から複合名詞内の語基への係り受けが生じる可能性がある．

複合語外の表現（α）から複合名詞（B）内への係り受けには，以下の二つのタイプがある．

1) α が B を連体修飾する場合：「α の B」（α は名詞*），埋め込み文（α）が B を連体修飾する場合など．
2) α（名詞*）と B が並列の場合：「α と B」「α や B」「α か B」など．

1) で埋め込み文（α）が B を連体修飾する場合については，α が B 全体に係る場合がほとんどであること，および埋め込み文の解析自体が日本語解析の大きな課題の一つになっていることから，本論文ではこれを対象外とし，1) では「α の B」のみを対象とする．一方，2) の場合，「α や B」「α か B」などは「α と B」と同様に扱うことができるため，「α と B」で代表させる．

ここでは，複合名詞を含む名詞句が「α の B」，「α と B」（α は名詞*，B は複合名詞）の場合について，複合語外（「α の」，「α と」）から複合名詞 B（語基列 $\beta_1 \cdots \beta_N$ より構成されているものとする）内の語基 β_k への係り受けを以下のように解析する．

①「α の B」の場合：

複合名詞 $\alpha\ \beta_i$ $(i=1 \sim N)$ において，α と β_i に係り受けが成立するような i の最小値 I を求める．I が求まれば「α の」の係り先は β_1 とする．I が求まらなければ「α の」係り先は複合名詞全体とする．

②「α と B」の場合：

複合名詞 $\alpha\ \beta_j$ $(j=1 \sim N)$ において，α と β_j に並列関係が成立するような j の最小値 J を求める．J が求まれば「α と」の係り先は β_J とする．J が求まらなければ「α と」の係り先は複合名詞全体とする．

複合語外から複合名詞内の語基への係り受けがある場合，係り元の名詞 α を B 内の係り先語基（β_I または β_J）の直前に挿入して B 内の語基とし，α と β_I または β_J で一つの部分複合語を生成し，既に生成されている B 内の部分複合語の構造内に組み込む．この

---

* 組織名は，地名，人名など他の固有名詞（部分複合語を含む）を含むより大きな構造をもつ部分複合語を生成することが多い．

** 役職は人名，組織名など複数の固有名詞（部分複合語を含む）と次々と結合して，より大きな構造をもつ部分複合語を生成することが多い．

*** 本接尾語の直前の語基，部分複合語が用言性名詞で，かつ当該用言性名詞の直前の語基，部分複合語と係り受けが成立する場合は，本ルールを適用せず，⑤で用言性名詞関連ルールを適用する．

* 複合名詞でもよい．

ように，複合語外の語 α が複合名詞 B 内の語基 β_k と係り受けがある場合，α と β_k の意味的関連を解析したうえで，α を B 内に組み込むことにより，複合名詞の訳出時に α と B を全く独立に訳出したうえで「α と B」や「α と B」などの訳出を行うのではなく，α と β_k の意味的関連を考慮し，α と B を一体化した柔軟な訳出が可能となる（その具体的方法は本論文の主たる課題でないのでここでは詳しく述べない）．**図3** に複合語外から複合名詞内への係り受けがある場合の複合名詞構造解析結果の変更例を示す．

なお，複合語外からの係り先が複合名詞全体の場合には，「α の B」や「α と B」において，α と B は独立に解析してよいため，上記のような変更を行わない．

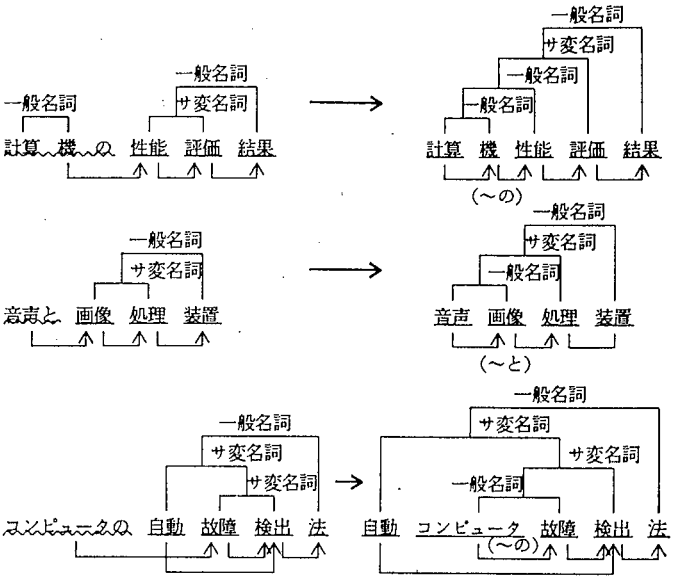


図 3　複合名詞構造解析結果の変更例（複合語外から複合名詞内の語基への係り受けがある場合）
Fig. 3　Modification of anlytical results for compound words.

4. 変換辞書の単語結合型辞書引き

解析辞書と変換辞書の単語収録単位は，2章で述べたように異なっており，変換辞書には語基のほかに，語基の組合せで構成される種々の長単位語を見出し語として収録する．従って，解析辞書と変換辞書の単語収録単位の違いを吸収する機構が必要となる．

ここでは，複数の語基より構成された複合語などの長単位語を変換辞書にある長単位語を含む見出し語の最適な組合せに再構成するため，複数の語基を組み合わせて変換辞書引きを行う単語結合型辞書引きを提案する．

4.1 複合名詞の単語結合型辞書引き

複合名詞 α は解析辞書に収録された語基の組合せとして，3章で述べた方法によって α の内部構造を解析し，部分複合語を生成する．また，3.3 節で述べたような名詞句中に出現する複合名詞（α）においては，α 外から α 内の語基への係り受け解析を行い，α 内語基への係り受けがある場合，係り元の名詞を α 内に組み込み構造解析結果を修正する．

α に対する変換辞書の単語結合型辞書引きは，α の部分複合語構造に基づき，以下のように行う．

①最大の部分複合語である複合語全体 α で辞書検索し，検索成功ならば☆終了，不成功ならば②へ．

②α を最も大きな部分構造（部分複合語や語基）の組合せ $\alpha_1 \cdots \alpha_N$ に分割する．

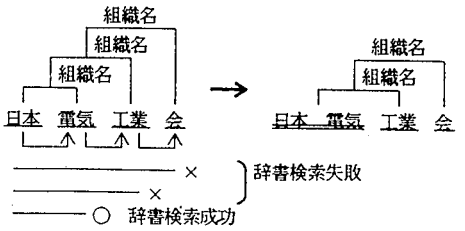


図 4　単語結合辞書引きによる複合語再構成の失敗例
Fig. 4　Errors of compound word reconstruction by combined word retrieval.

③$\alpha_1 \cdots \alpha_N$ において α_i が部分複合語ならば，α_i で辞書検索する．辞書検索成功ならば，α_i に包含される部分複合語があってもこれ以上小さな単位で辞書検索しない．辞書検索不成功ならば☆，α_i を②と同様にして最も大きな部分構造の組合せ $\alpha_1' \cdots \alpha_M'$ に分割し，$\alpha_1' \cdots \alpha_M'$ において α_j' が部分複合語ならば③と同様に α_j' で辞書検索する．以下，同様の処理を繰り返し，辞書検索すべき部分複合語がなくなった時点で処理を終了する．

上記，①～③の処理によって抽出された辞書検索成功した部分複合語を用いて α を再構成したものが，α の単語結合型辞書引きの結果である．図2に単語結合型辞書引きの例を示す．

☆ 部分複合語の品詞と検索された語の品詞が一致しなければならない．例）完全自動化→サ変動詞名詞

☆ 部分複合語の品詞が固有名詞（組織名）の場合には，**図4** の例に示すような誤りを防ぐため，辞書検索が失敗してもより小さな部分複合語による辞書検索を行わない．

4.2 複合用言の単語結合型辞書引き

サ変動詞語幹 α，形容動詞語幹 β となる複合名詞に用語の活用語尾・付層語 (γ) が連続すると，以下のように $\alpha\gamma$ や $\beta\gamma$ の形の複合用言を構成する．複合用言を構成する場合，α や β の内部構造は，複合名詞ほど複雑にはならない．

① γ=「する」の場合（英語の動詞に対応）

例）故障 検出 する，
工業 化 する

② γ=「だ」，「な」，「の」の場合（英語の形容詞に対応）

例）栄養 豊富 だ，近代 的 な
小型 軽量 の，全 自動 の

③ γ=「に」（英語の副詞に対応）

例）超 高速 に，効率 的 に

ここで，α，β については，3章の方法を適用して α，β の内部構造を解析し，必要に応じ α，β からの，α または β 内の語基への係り受けを解析しておく．

α，β に対する変換辞書の単語結合辞書引きは，以下のようにして行う．

1) α，β 外から α，β 内の語基への係り受けがない場合

$\alpha\gamma'$，$\beta\gamma'$ で単語結合型辞書引きを行う．ここで γ' は γ=「な」の場合，γ'=「だ」とする以外，γ と同じ字面を設定する．検索成功ならば終了，不成功ならば 2) と同様の処理を行う．

2) α，β 外からの α，β 内の語基への係り受けがある場合

α，β の構造解析結果に基づき α，β 内の語基に適当な付属語などを補い，通常の文の形に変換し，変換処理を行う．

例）故障検出する→故障 を 検出する
コンピュータの 故障検出する
→ コンピュータの故障 を 検出する
栄養豊富な 食品 → 栄養 が 豊富な食品
研究熱心な 人 → 研究 に 熱心な人

5. 有効性の評価

5.1 複合名詞の構造解析法の評価

3章で提案した複合名詞の構造解析法の有効性を確認するため，情報通信，AI 関連の新聞記事リード文 200 文に出現する複合語 201 件のうち，未知語を含まないもの 167 件について構造解析を行った．

その結果，全体の 94.6% に当たる 158 件が正しく解析できた．構造解析失敗例を**図 5** き示す．構造解析に失敗した 9 件のうち 7 件は，図 5 の例 3 のように一般語が数語結合して固有名詞（商品名，企業名など）を構成するものであり，ある部分複合語（名詞句相当）が後続する固有名詞を説明的に修飾する同格表現である．複合名詞の構造解析法の評価に用いた複合名詞（167 件）に含まれる語基数は 568 で，複合名詞当たりの平均語基数は約 3.4 である．構造解析に失敗した複合名詞（9 件）に含まれる語基数は 46 で，複合名詞当たりの平均語基数は約 5.1 となる．このことから，含有語基数の多い複合名詞は，含有語基数の少ない複合名詞に比べ，構造解析に失敗しやすいことが実験結果から裏付けられた．

今後，解析精度を向上させるには，複合名詞に未知語を含む場合の扱い，同格表現の解析などについて検討を進める必要がある．

5.2 単語結合型辞書引きの効果

本方式を日英機械翻訳システム ALT-J/E[8] に組み込みその効果を検証した．ALT-J/E への適用結果によれば，日本語の一般語を解析辞書に収録した場合，長単位語も収録すると収録語数は約 13 万語となるが，短単位語のみ収録すると収録語数は半分程度の約 7 万語となった．なお，語数が多く大部分が複合語である専門用語についてはほとんど解析辞書に収録する必要

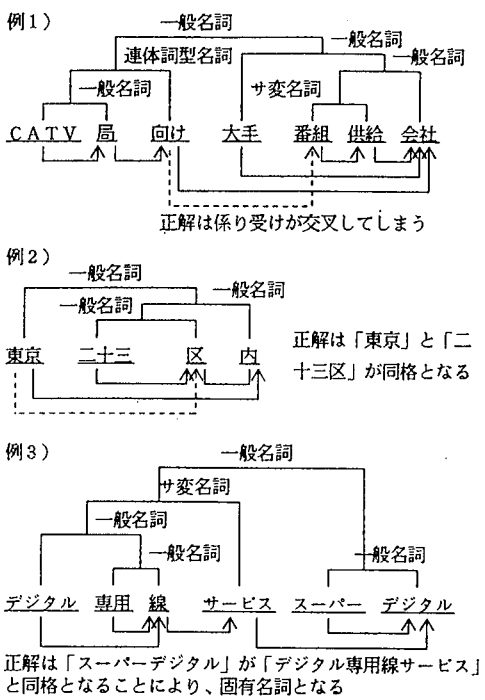


図 5 複合名詞の構造解析失敗の例
Fig. 5 Errors of compound word analysis.

がないので，解析辞書のコンパクト化の効果はさらに顕著となった．

また，解析に必要な意味属性などの意味情報を多くの複合語に対して付与しなくてすむこと，解析辞書と変換辞書の独立性を確保できることなどにより大規模辞書の効率的な構築・維持管理が可能となった．

6. あとがき

機械翻訳システムにおいて，解析辞書（日本語辞書）には原則として語基のみ収録し，変換辞書（対訳辞書）には目的言語に応じて適切な訳出を行うため語基のほかに複合語などの長単位語も収録することにより，解析辞書と変換辞書の独立性を確保し，辞書のコンパクト化が可能な辞書システムが構成できることを論じた．

さらにこのような辞書構成法を可能とするため，各辞書の単語収録条件の差を吸収することを狙い，複合語を解析辞書にある語基の組合せとして内部構造を解析し，その結果に基づき複合語内の語基を組み合わせて変換辞書引きを行う単語結合型辞書引きを提案した．

本方式は日英機械翻訳システム ALT-J/E に組み込まれ，その効果が確認されている．ALT-J/E への適用結果によれば，本手法は，大規模辞書を効率的に構築し，維持管理することを可能とした．特に，解析に必要な意味属性などの意味情報を多くの複合語に対して付与しなくてすむ点などに利点がある．

また，単語結合型辞書引きの前処理として行われる複合名詞の構造解析においては，部分複合語の概念を新たに導入し，複合語外から複合名詞内の語基への係り受けも考慮して，語基間の係り受け解析結果を基に，複合名詞内の語基がどのように結合して部分複合語を構成し，それらがより大きな部分複合語にまとめられるかという点を自動解析する方法を提案し，その有効性を示した．

今後，複合名詞に未知語を含む場合の扱い，複合名詞がもっと複雑な名詞句中にある場合の扱い，係り受け解析と部分複合語生成の融合等について検討を進める予定である．

謝辞 本論文の方式は日英機械翻訳システム ALT-J/E を構成する技術の一つとして開発してきたものである．本システムの開発を通じて本方式に関する有意義な議論を頂いた白井主任研究員，中岩研究主任をはじめ，翻訳研究グループの諸氏に感謝する．また，複合名詞の構造解析の評価でご協力を頂いた NTT ソフトウェアの植田久和氏に感謝する．

参考文献

1) 宮崎正弘：日本文音声変換のための数詞読み規則，情報処理学会論文誌，Vol. 25, No. 6, pp. 1035-1043 (1984).
2) 宮崎正弘：係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Vol. 25, No. 6, pp. 970-979 (1984).
3) 野村雅昭：複次結合語の構造，国語研報告，Vol. 49, pp. 72-93 (1973).
4) 田中，水谷，吉田：語と語の関係について，情報処理学会自然言語処理研究会資料，41-4 (1984).
5) 西野，藤崎：漢字複合語の確率的構造解析，情報処理学会論文誌，Vol. 29, No. 11, pp. 1034-1042 (1988).
6) 藤田，辻井，長尾：漢字連続複合語の解析，第28回情報処理学会全国大会論文集，7M-3 (1984).
7) 石崎雅人：日本語複合名詞の解析，第35回情報処理学会全国大会論文集，1T-1 (1987).
8) 池原，宮崎，白井，林：言語における話者の認識と多段翻訳方式，情報処理学会論文誌，Vol. 28, No. 12, pp. 1269-1279 (1987).

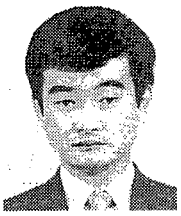
(平成4年3月 6 日受付)
(平成5年1月 18日採録)

宮崎 正弘（正会員）

昭和21年生．昭和44年東京工業大学工学部電気工学科卒業．同年日本電信電話公社に入社．平成元年より新潟大学工学部情報工学科教授．この間，大型情報処理装置 DIPS の開発，計算機システムの性能評価法の研究，日本文音声出力システムや機械翻訳などの自然言語処理の研究に従事．現在，自然言語理解，機械翻訳などの研究に従事．工学博士．電子情報通信学会，人工知能学会各会員．

池原 悟（正会員）

昭和19年生．昭和42年大阪大学基礎工学部電気工学科卒業．昭和44年同大学大学院修士課程修了．昭和44年日本電信電話公社に入社．現在，NTT 情報通信網研究所知識処理研究部勤務．この間，数式処理，トラヒック理論，自然言語処理の研究に従事．工学博士．昭和57年情報処理学会論文賞受賞．電子情報通信学会，人工知能学会各会員．



横尾　昭男（正会員）

昭和32年生．昭和55年電気通信大学電気通信学部電子計算機学科卒業．昭和57年同大学院電子計算機学専攻修士課程修了．昭和57年日本電信電話公社に入社．現在，NTT情報通信網研究所知識処理研究部勤務．この間，自然言語処理の研究に従事．現在，日英機械翻訳システムにおける日英構造変換処理や翻訳辞書の研究に従事．電子情報通信学会，人工知能学会各会員．